DOMAIN-SPECIFIC APPROACHES TO SENTIMENT ANALYSIS IN ECONOMICS

Débora Cristina Livino de Oliveira

Dissertation
Master in Data Analytics

Supervised by
Prof. João Cordeiro
Prof. Pavel Brazdil

2018

Acknowledgments

Firstly, I would like to thank my supervisors, Professor João Cordeiro and Professor Pavel Brazdil, for their time, guidance and support throughout this project. I am especially grateful to Professor Pavel Brazdil for his availability, understanding, incentive and valuable insights.

To the researchers involved in this project for their work.

To my parents, Patricia and Edmilson, for the support and encouragement since ever. I am blessed to be their daughter.

Finally, to my husband, Roberto. This work would be unattainable without his support. Thank you very much for being such an amazing partner to me.

Abstract

Sentiment analysis evaluates people's declarations from written language aiming to automatically determine opinions, sentiments, evaluations, appraisals and emotions, identifying whether the text expresses a positive, neutral or negative perspective.

This thesis performs the sentiment analysis in opinion articles in the specific domain of economics, exploring linguistics and computer science techniques and analyzing the effects of adding domain-specific terms to a given lexicon.

Our approach for sentiment analysis in European Portuguese is based on a combination of a general-purpose lexicon, SentiLex-PT02, and a domain specific lexicon, EconoLex-PT, constructed manually by experts.

The goal is to verify whether the combination of a general-purpose lexicon and a domain-specific lexicon achieves better performance than just a general-purpose lexicon on its own.

Keywords: Sentiment Analysis, Lexicon-based approach, Domain-Specific, Economics

Resumo

Análise de Sentimento avalia declarações de pessoas expressas na linguagem escrita com o propósito de automaticamente determinar opiniões, sentimentos, avaliações, apreciações e emoções, identificando se o texto expressa uma perspectiva positva neutra ou negativa.

Esta tese realiza análise de sentimento sobre artigos de opinião especificamente no domínio de Economia, explorando técnicas do domínio de linguísticas e de ciência da computação e analisando os efeitos de se adicionar termos de domínio específico a um dado léxico.

Nossa abordagem para análise de sentiment em Português Europeu é baseada emu ma combinação de um léxico de propósito geral, SentiLex-PT02, e um léxico de domínio específico, EconoLex-PT, construído manualmente por especialistas.

O objetivo é verificar se a combinação de um léxico de propósito geral e um léxico de domínio específico alcança melhor performance que apenas um léxico de propósito geral por si só.

Palavras-chave: Análise de Sentimento, Abordagem com base em Léxico, Domínio Específico, Economia

Contents

Index of Figures

Index of Tables

1. Introduction

1.1 Motivation

Our perception of facts, people and things are highly influenced by the way others see the world. Similarly, the opinion of others is relevant in the decision-making process. Sentiment analysis (SA) aims to automatically determine opinions, sentiments, evaluations, appraisals and emotions, identifying whether a text is subjective or not and, if subjective, whether it expresses a positive, neutral or negative perspective.

The internet has become one of the most important sources of information for all users, from individuals to organizations. Companies are interested in the feedback from customers as well as in news about competitors and suppliers. From the consumer perspective, consulting other customer reviews before purchasing is a common behaviour.

The production and consumption of data increases at an exponential rate and highlights the relevance to convert raw unstructured data into valuable information. In this context, Sentiment Analysis emerged and has received increased attention over the years due to the significant role it plays in the decision-making process and its wide application in several fields, such as business, politics and economics.

The complexity of the economic systems has become more evidence with the advent of the internet. Everyone can publish content and this access to communication tools has profoundly changed the power of communication itself and the ability to influence others. It affects the way people discover, read and share news, ideas and opinions. Therefore, tracking and monitoring the content that has been published is extremely important due to the impact of this new dynamic to the economic systems.

The motivation for this work comes from the desire to evaluate the performance of a domain-specific lexicon and provide a tool to perform sentiment analysis in Economics for the European Portuguese since the resources available for this language are rather scarce.

1.2 Problem to be studied

This Master thesis consists of the sentiment analysis in opinion articles in the specific domain of economics, exploring linguistics and computer science techniques. The goal is to investigate aspects of lexical and sentence semantics related to the sentiment expression in European Portuguese.

This work has been developed in collaboration with an R&D group, which involves researchers from the School of Economics and Management of the University of Porto (FEP), Faculty of Arts and Humanities of University of Porto (FLUP) and University Beira Interior (UBI).

As some researchers have shown (Forte and Brazdil, 2016; Almatarneh and Gamallo, 2017), domain specific lexicon may be very useful in this process since it can enhance the performance. In this project, the effects of adding domain-specific terms to a given lexicon as well as the extent to which the results can be boosted will be evaluated.

Our approach for sentiment analysis in European Portuguese is based on a combination of a general-purpose lexicon, SentiLex-PT02, and a domain-specific lexicon, EconoLex-PT, constructed manually by experts.

The goal is to verify whether the combination of EconoLex-PT and SentiLex-PT achieves better performance than just SentiLex-PT on its own. The domain specific lexicon includes multi-words, adjective, noun and verb phrases.

1.3 Thesis Structure

The overall thesis is structured as follows:

Chapter 2 presents an overview of sentiment analysis. It presents the main concepts involving text in this context, text processing methodologies, text representations, pre-processing techniques, the model validation, evaluation measures, levels of analysis and main approaches.

Chapter 3 describes the methodology developed in this thesis. The algorithm developed to perform all analysis, the corpus and the pre-processing techniques are described.

Chapter 4 shows our case study, presenting the results of sentiment analysis for a series of experiments.

Chapter 5 presents the main conclusions and the limitations of this work, besides suggest future works that could be done to improve our results.

2. Sentiment Analysis Overview

Sentiment analysis evaluates people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions from written language towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes (Liu, 2012). The same author declares that the textual information can be categorized into two types: facts and opinions.

Facts are objective expressions about entities, events and their properties while opinions are usually subjective expressions that describe people's sentiments toward the same objects.

Sentiment Analysis is a field of research in Text Mining (TM). Feldman and Sanger (2007) explains that TM extracts useful information from unstructured textual data sources by identifying and exploiting interesting patterns through techniques and methodologies from the areas of information retrieval, information extraction, and corpus-based computational linguistics.

According to Esuli and Sebastiani (2006), the following subtasks of Sentiment Analysis can be identified:

- determining whether a given text has an objective or subjective nature, respectively representing a fact or expressing an opinion;
- establishing the polarity of a given subjective text, expressing a positive or a negative opinion;
- deciding the strength of the polarity of a given subjective text, expressing a weakly, mildly, or strongly positive or negative opinion; and
- extracting opinions from a text.

Batrinca and Treleaven (2014) states that sentiment analysis applies natural language processing, computational linguistics and text analytics in order to identify and extract subjective information in source materials.

2.1 Text processing methodologies

Aiming to perform the text pattern analysis and mining in the plain text, the following methodologies can be applied:

- N**atural Language Processing**

  Natural Language Processing (NLP) is an automatic method that manipulates and processes the natural language, enabling machines to interpret the human language and being to handle take human-produced text as input as well as produce natural looking text as outputs (Goldberg, 2017).

- **Information Extraction**

  Information Extraction (IE) extracts the meaningful elements from large amount of text. IE analyzes unstructured text by identifying key phrases and relationships within text and transforming a corpus of textual documents database into a more structured database (Gaikward et al., 2014).

- **Information Retrieval**

  Information Retrieval (IR) extracts relevant and associated patterns based on a set of words or phrases. IR finds unstructured data that matches the requirements within large collections, for instance, supporting users in browsing or filtering document collections or further processing a set of retrieved documents (Manning et al., 2008).

Besides the methods presented above, there are others, such as: **Clustering**, in which similar terms or patterns are extracted and grouped from many documents; **Text Summarization**, which produces a concise representation of a large document or a collection of documents; **Categorization**, which identifies the main subject of a document by putting it with a given set of topics etc.

2.2 Text representation

Documents should generally be converted from the original format into a more manageable representation since the common classifiers and learning algorithms cannot directly process the text documents (Feldman and Sanger, 2007).

- **Bag-of-Words**

  Bag-of-Words (BoW) exploits all words aiming to categorize documents by analysing and classifying different bags of words. It is worth to mention that BoW ignores the

order in which the words appear. The goal of this method is to identify which bag a certain piece of text comes from by matching the different categories.

- **Document Term Matrix**

  Document Term Matrix (DTM) represents the frequency of terms in a corpus in which displaying documents as rows, terms as columns and the frequency of terms as the entries.

- **Term Frequency-Inverse Document Frequency**

  Term Frequency-Inverse Document Frequency (TF-IDF) identifies the most relevant words to a text document in a corpus. TF-IDF scores the importance of terms in a document based on often they are present across multiple documents. In other words, if a term appears frequently in a document, it is important and, consequently, the word should have a high score. On the other hand, if a word appears in many documents, it is not an effective identifier, so the word should receive a low score.

2.3 Text pre-processing techniques

Data pre-processing consists of cleaning the text by removing its uninformative parts. This task aims to reduce the computational complexity of the classification process by reducing the data dimensionality and improving the accuracy of the sentiment analysis while enhancing the quality of the data.

Some of the most basic pre-processing tasks involve the removal of punctuations and numbers as well as the conversion of the characters from uppercase to lowercase. Other relevant transformations on the data that will be performed in this work are detailed below:

- **Tokenization**

  Tokenization splits sequences of characters into smaller and more meaningful parts called tokens. The breakdown may occur at different levels (e.g.: paragraphs, sentences, words and even syllables or phonemes) and the main challenge is to distinguish between a period that indicates the end of a sentence and a period that is

part of a previous token in identifying sentence boundaries (Feldman and Sanger, 2007).

- **Stopwords removal**

  Stopwords are words that do not add relevant information to the sentiment analysis process for presenting a non-predictive and non-discriminating content. The removal of these words is advantageous for reducing the dimensionality of term space and, consequently, improving the retrieval rate and boosting the prediction results.

  Stopwords can be grouped into two categories: general and domain-specific. The general group includes the standard stopwords available in the public domain as well as the non-standard ones generated inside information retrieval or text categorization systems, and the domain-specific group contains the words that do not present a discriminant value within a specific domain or context (Makrehchi and Kamel, 2008).

- **Dealing with negation**

  Negation words shift the sentiment expressed in sentences. However, handling these terms is a complex task since not always there is a direct negative form in which the negation and the negated word are neighbours. For instance, negation can have as target a verb, a subject, or an adjective or adverbial phrase. Furthermore, these words that express the opposite sentiment can also enhance the positive sentiment of a sentence, instead of reversing it.

- **Part-of-Speech tagging**

  Part-of-Speech (POS) tags split words into grammatical categories according to the function that each of them plays in the sentence, providing information about the semantic content of a word (Feldman and Sanger, 2007). POS correspond to the lowest level of syntactic analysis and include nouns, pronouns, adjectives, verbs, adverbs, conjunctions, prepositions and interjections.

- **Stemming and Lemmatization**

  Stemming and lemmatization have a similar goal: normalize words. However, they adopt different approaches and levels of complexity and accuracy. Stemming removes derivational suffixes and inflections to reduce all words with the same stem

to a common form while lemmatization removes inflectional ending and returns the dictionary form of a word by performing vocabulary and morphological analysis (Balakrishnan et al., 2014). The lemmatization process requires the understanding of the context, the specification of the POS of each word in the sentences and then finally the definition of the lemma (Jivani, 2011). Lemmatization tends to be a more accurate method, but it presents a more complex process.

- **Multi-words**

  Choueka (1988) states that a multi-word expression is a sequence of neighbouring words whose meaning or connotation cannot be derived from the meaning or connotation of its components. Following this statement, Schone & Jurafsky (2001) developed three criteria for classifying an expression as a multiword, which are: semantic non-compositionality, lexical non-substitutability and syntactic non-modifiability.

  Non-compositionality implies that the expression cannot be deduced from the meaning from the meaning of the individual words. Non-substitutability indicates that substituting a word with a synonym will change the original meaning of the expression. Non-modifiability means that the structure of the expression cannot be modified without changing its original content.

  Nevertheless, according to **Gouws** et. al (2013) not every multi-word presents all three attributes simultaneously and to the same extent, even if the semantic non-compositionality can be noticed in most cases.

2.4 Evaluation measures in Information Retrieval

Evaluation measures usually reveal which classes are difficult for the classifier, either because of lack of training data, or poor quality of the data, or simply because some classes are hard to identify even for humans (Farzindar and Inkpen, 2018).

- **Accuracy** is the number of correct predictions over the total number of predictions.
- **Precision** is the number of correct predictions over the total number of true positives and true negatives. For instance, a high precision means that the majority of items classified as positive indeed belongs to the class positive.

- **Recall**, also known as sensitivity, is the number of correct predictions over the total number of true positives and false negatives. A high recall means that the majority of the positive items were labelled as belonging to the class positive.

- **F-Measure** combines Recall and Precision into a single metric, represented by the harmonic mean of both, which receive equal weights.

However, the intensity of the sentiments in this project will have seven levels, on a scale from -3 to +3. The metric should be able to compare the predicted values against the actual values and capture the extension of the errors. Therefore, the evaluation measure selected to evaluate the performance is the Mean Absolute Error.

- **Mean Absolute Error (MAE)** measures the average magnitude of the errors in a set of predictions with regression models. MAE is the average over the test sample of the absolute differences between forecast and actual observation in which all individual differences have equal weight.

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n} | y_j - \hat{y}_j | \quad (1)$$

2.5 Model validation

The model validation adopted to assess the accuracy of the predictive model was the Leave One Out Cross Validation (LOOCV). Cross validation aims to evaluate the capability of the model to generalize to an independent dataset and make predictions for new observations.

Leave One Out is a procedure that uses a single observation for validation while the remaining observations from the sample are used as the training data. This procedure is repeated such that each observation is used once as the validation data. In other words, the data is trained on all observations except for one and a prediction is made for that one observation.

2.6 Levels of Sentiment Analysis

Sentiment analysis can be performed at three main levels: document level, sentence level, and entity and aspect level.

- **Document level**

Document level sentiment classification aims to evaluate documents based on the overall opinion contained in the document and classify whether it expresses a positive or negative sentiment (Pang et al., 2002; Turney, 2002). At this level, it is assumed that each document expresses opinion on only one topic or entity (Liu, 2012). However, a single document may contain multiples and divergent opinions and this is why analysis at the sentence level may me more accurate than at the document level.

- **Sentence level**

Sentence level sentiment classification aims to evaluate documents based on the opinion contained in each sentence and classify whether it expresses a positive, negative, or neutral sentiment (Liu, 2012). At this level, it may be relevant to determine whether the sentence expresses a sentiment (subjective sentence) or not (objective sentence) (Wiebe et al. 1999). For subjective sentences, there is a third classification called neutral for sentences that do not express sentiment (Wilson et al., 2004).

Sentence level assumes that the whole sentence expresses a single opinion from a single opinion holder. However, Liu (2010) pointed out that a single sentence may express more than one opinion for compound sentences as well as Wilson et al. (2004) emphasized that not only a single sentence may present different opinions, but it may also present subjective and objective clauses.

- **Entity and aspect level**

Considering that an opinion consists of a sentiment (positive or negative) and a target (of opinion), aspect level directly evaluates the opinion itself instead of evaluating the language constructs (Liu, 2012). Still according to the author, there following four approaches can be performed to extract the explicit aspects: frequent nouns and noun phrases; opinion and target relations; topic models; and supervised learning methods.

2.7 Approaches of Sentiment Analysis

There are two main approaches for the problem of Sentiment Analysis: lexicon-based and machine learning-based. According to Turney (2002), the lexicon-based approach

computes the orientation for a text according to the polarity and strength of words, phrases or texts contained in the document. As stated in Pang et al. (2002), the machine learning-based approach build a classifier based on labelled instances of texts or instances.

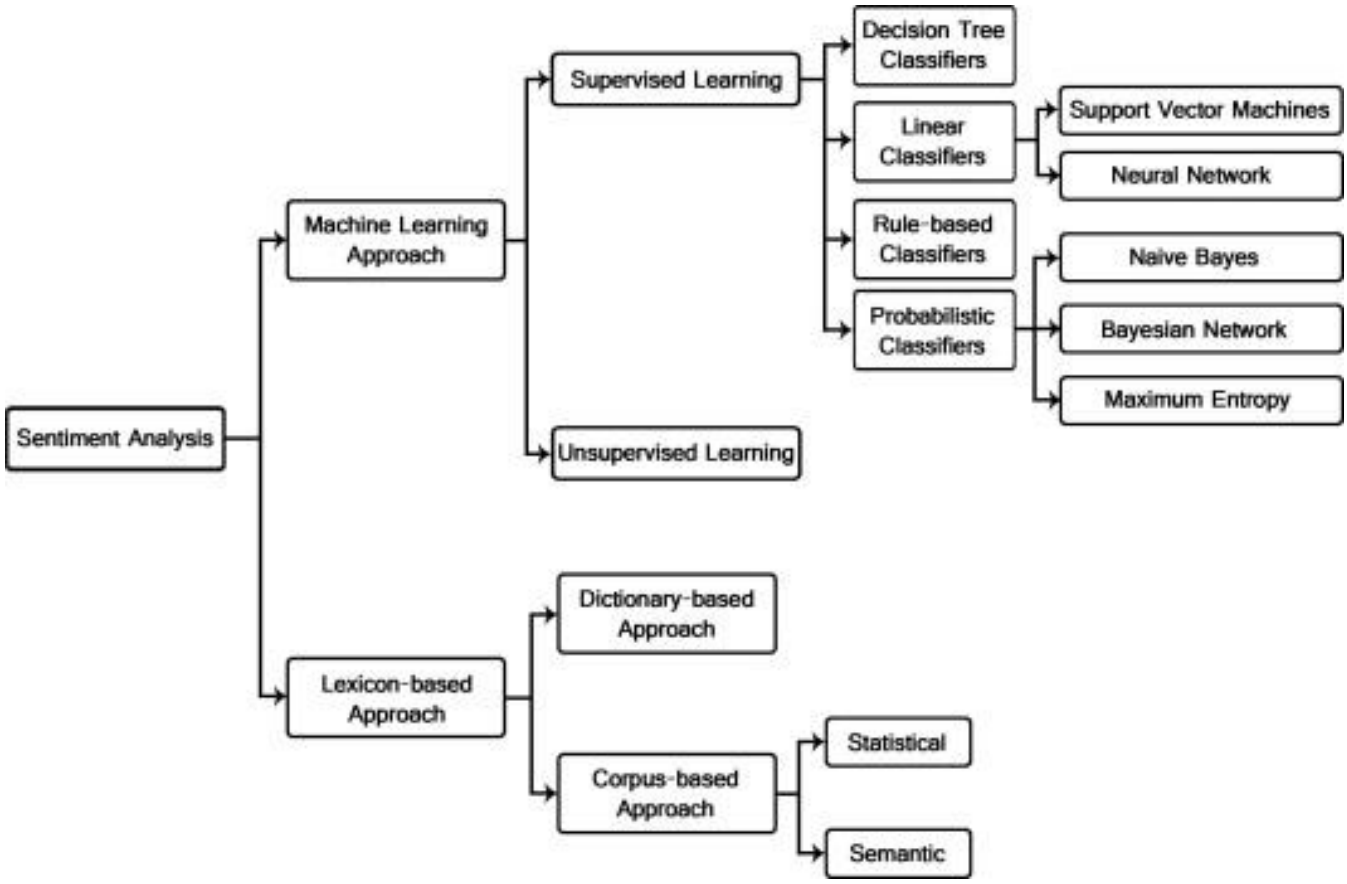

Figure 1 Sentiment Classification Techniques (Walaa Medhat et al., 2014)

2.7.1 Lexicon-based approach

In the lexicon-based approach, the sentiment of a text is derived from the sentiment aggregation of its composing words, usually represented as a bag-of-words (Boiy and Moens, 2009). The sentiment value of each word or multi-word is assigned to a dictionary by experts and the aggregation is calculated either by the average or the sum of the individual value of the word.

Taboada et al. (2011) mention that the lexicon-based method is robust across different domains and texts. Furthermore, Forte and Brazdil (2016) evaluated that a domain-specific lexicon enhances the performance.

There are three types of lexicon-based approach: manual, dictionary-based and corpus-based.

- **Manual approach**

In the manual approach, the dictionary is collected and built manually, word by word, which is an inefficient process. For being a very time-consuming activity, this method rarely is used except when in combination with an automatic method.

- **Dictionary-based approach**

Originally used in sentence and aspect level of analysis (Hu and Liu, 2004), the dictionary-based approach determines the word sentiment based on a set of seed opinion words that will be increased by adding lists of synonyms or antonyms words (Fellbaum, 1998; Ding et al., 2008).

The process of scanning and adding new words to the seed list continues while new words are being found and, at the end, a manual inspection is performed in order to remove errors (Liu, 2011; Edelman and Ostrovsky, 2007; Huang et al., 2007).

The main advantage of the dictionary-based approach is the efficiency with which the sentiment words can be processed. However, this method is not able to distinguish opinion words that present different meanings in domain and context-specific orientations (Liu, 2011).

- **Corpus-based approach**

The corpus-based is the more suitable approach for domain specific lexicon since it addresses the problem of interpreting the meaning of words that vary depending on the context. Ding et al., 2008 states the corpus-based method evaluates co-occurrence patterns of words in order to determine the sentiments of words or phrases. Liu and Zhang (2012) complement the statement adding that this method also rely on syntactic and on seed list of opinion words to find more opinion words in a large corpus.

Liu (2012) point out the corpus-based approach has been widely used in the following situations: aiming to reveal new sentiment words and their respective semantic orientation from a domain corpus over a seed list of general-purpose sentiment words; and for adjusting a general-purpose sentiment lexicon to a domain-specific corpus for sentiment analysis.

2.7.2 Machine Learning-based approach

Machine learning algorithms can perform sentiment analysis through regular text classification while making use of syntactic and linguistic features (Medhat et al., 2014). In this approach, which is fundamentally a supervised learning process, the classifier is built based on a training set aiming to identify the sentiment of new texts according to certain patterns.

Supervised learning problems are grouped into classification and regression problems. The difference between them is that the output is a category for classification problems while the output is a real value for regression problems. Random forest is an example of one of many algorithms able to evaluate supervised learning problems and it is able to solve both classification and regression problems.

In spite of being focused on analyze the performance of a domain-specific approach, this project intends to compare the results to an alternative machine-learning based approach to be defined.

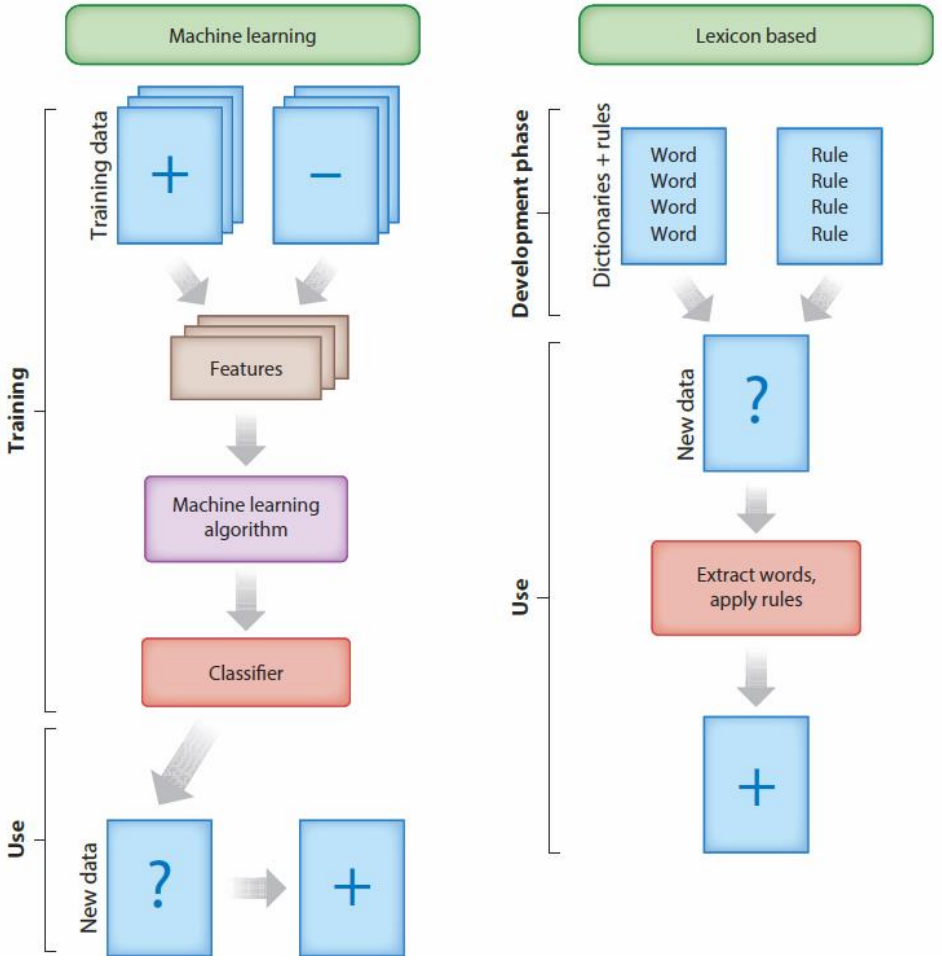

Figure 2 Machine Learning and Lexicon-Based approaches to Sentiment Analysis (Taboada, 2016)

3. Methodology

This chapter describes the methodology that have been used in this thesis. It starts by presenting the dataset and lexicon as well as the software that have been used in this work. Further on, it describes the Sentiment analysis algorithm and its key points.

3.1 Dataset and Lexicon

- **Dataset**

The object of this study consists of a collection of texts from the economic domain, written in European Portuguese that contain news in Economics. There is a total of 45 articles of opinion extracted from digital newspapers that were previously analysed by researchers from the Faculty of Arts and Humanities of University of Porto in order to identify linguistic elements that contain sentimental content.

The experts from the linguist field evaluated the contribution of the sentiment lexicon to the expression of a positive, neutral or negative opinion. For each one of the 370 sentences considered relevant to the 45 texts, the overall sentiment value was manually determined on a discrete scale, ranging from -3 (strongly negative) to +3 (strongly positive). The same scale was used to determine the sentiment value of each word (noun, adjective, adverb or verb), multi-word or short phrase appearing in the lexicon EconoLex-PT.

The sentiment value of new texts will be determined in the usual way, by summing up the sentiment values of all terms that appear in the lexicon. The resulting value is then re-scaled, so that the final value would be in the interval from -3 to 3.

- **SentiLex-PT02 Lexicon**

SentiLex-PT02 is a general-purpose sentiment lexicon made up of 7,014 terms in the lemma form and 82,347 inflected forms. It was designed by Silva, Carvalho and Sarmento for the extraction of sentiment and opinion about human entities in texts written in Portuguese.

- **EconoLex-PT Lexicon**

EconoLex-PT is a domain-specific lexicon whose goal is to identify linguistic elements to extract sentiment and opinion from texts of the economic domain written in European Portuguese.

This domain-specific lexicon is an outcome of the analysis over the 45 articles of opinion from the economic domain and developed by the researchers from the Faculty of Arts and Humanities of University of Porto.

3.2 Software and Programming Language

R, which is a programming language and environment for statistical computing and graphics, was the software used to develop all the analysis regarding this thesis (version 3.5.1). It is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form and was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand and

R has several text mining packages available to perform all the tasks involving the sentiment analysis, such as: tm, tokenizers, tidyverse, dplyr, plyr and ptstem. Besides these packages, xlsx was also used to export the results from the R software to the Excel software.

3.3 Sentiment Analysis algorithm

The author of this thesis developed an algorithm in R to execute the tasks need to the answer to the following question: "Does the combination of EconoLex-PT and SentiLex-PT achieve better performance than just SentiLex-PT on its own?"

- **Algorithm description**
  - Read the corpus;
  - Read the EconoLex;
  - Read the SentiLex-PT02;
  - Build a third lexicon that merges EconoLex and SentiLex;
  - Texts pre-processing;

- Establish the priority sequencing, in which EconoLex terms have higher priority than SentiLex-PT02 terms as the first rule and that the higher the number of words in a term the higher the priority as the second rule;
- Ensure that each term is computed only once, in case a term of the text is present in the lexicon as a simple individual word and also as a multi-word;
- N-gram definitions;
- Error calculation;
- MAE calculation;
- Leave One Out Cross Validation;
- Error calculation;
- MAE calculation.
- Wilcoxon test for paired samples;
- Performance improvement adding synonyms.

3.4 Pre-processing techniques

In order to prepare the data for being processed by the algorithm, the following pre-processing techniques were applied to this work:

- Removal of irrelevant information;
- Removal of numbers and unnecessary spaces and punctuation;
- Conversion to lower case;
- Removal of stopwords.

4. Case Study Results

This thesis aims to evaluate the contribution of a domain-specific lexicon plays in a sentiment analysis system. In particular, our goal is to verify whether the combination of a general-purpose lexicon together with a domain specific lexicon achieves better performance than just a general lexicon on its own. The method involves three steps briefly summarized below.

In the first step, the usual method of calculating the sentiment polarity is followed. Each term of the text found in the lexicon has its respective score assigned and then the summation of all term scores is calculated at the phrase level. In other words, the sentiment polarity of each phrase is defined by the sum of polarity of terms that compose it.

In the second step, an adjustment a factor is applied in order to perform an adjustment needed due to a difference of scales between the predicted sentiment value and the sentiment value attributed by domain experts.

In the third step, the Leave One Out Cross Validation procedure is performed to evaluate the accuracy of the predictive model.

Before discussing each step in more detail, we give an overview of the lexica used.

4.1 Lexicon Overview

The three steps of analysis were applied to perform the sentiment analysis for each phrase with each one of the following lexica:

- General-purpose SentiLex-PT
- Specific-domain EconoLex-PT
  Combination of SentiLex-PT and EconoLex-PT, prioritizing the domain-specific entries over the general-purpose entries.

4.1.1 SentiLex-PT02 Overview

SentiLex-PT02 ranges from -1 to +1. The sentiment is represented as -1 when negative, as 0 when neutral and as +1 when positive. In this general-purpose lexicon, the sentiment

entries are the terms, which may be single words or multi-words units, and the corresponding polarity assignment. In the form of *lemma*, most of terms belong to the class of the adjectives. In the flexed form, there are more verbs than adjectives, but only because are more forms of verbal inflection than adjective inflection (see **Table 1**).

Class	Lemma	Flex
Noun	1,080	1,280
Adjective	4,779	16,863
Verb	489	29,504
Multi-word	666	34,700
Total	**7,014**	**82,347**

Table 1 SentiLex-PT02 grammatical classes

SentiLex-PT02 is mostly comprised of terms whose sentiment is negative. Evaluating the 82,347 entries contained in this lexicon, 66% of them present a negative sentiment value against only 25% of the terms whose sentiment is positive, while the remaining 9% of terms are classified as neutral (see Table 2).

Class	Sentiment value
Negative	66%
Neutral	9%
Positive	25%
Total	**100%**

Table 2 SentiLex-PT02 Sentiment polarity distribution

From this moment on, this lexicon will be called only SentiLex-PT, omitting the allusion of its version.

For demonstration of SentiLex-PT02 application, following a phrase extracted from the text called "Declínio", written by Luís Todo Bom and published by Jornal Expresso:

"A criação de valor é uma constante neste processo de boa gestão empresarial em que se constroem unidades empresariais sólidas, bem dimensionadas, tecnologicamente evoluídas, capazes de jogarem o jogo da globalização e com uma boa performance bolsista."

The following three terms present in the text above were found in the SentiLex-PT02: "boa performance" [+1], "capazes" [+1] and "constante" [+1], respectively a multi-word unit and two adjectives.

4.1.2 EconoLex-PT Overview

EconoLex-PT is a domain-specific lexicon that is made up 1,246 terms in the form of l*emmas* and 2,811 in the flexed form, as well as its sentiment scores ranging from -3 (strongly negative) to +3 (strongly positive).

In this lexicon, the sentiment polarity between positive, neutral and negative terms are more balanced than in the SentiLex-PT: 46% are positive, 35% are negative and 19% are neutral.

It can be noted that the domain-specific lexicon is predominantly positive, in opposition to the general-purpose lexicon.

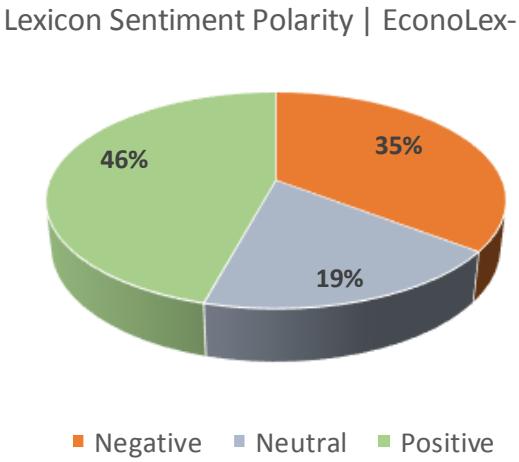

Figure 3 Sentiment polarity histogram for EconoLex-PT Lexicon

Figure 4 shows the distribution of the terms present in the EconoLex-PT between each one of the seven classes that represent the sentiment value in its polarity scale. The

polarity class that stands for the highest number of terms is the +1, with 28% of the total terms, followed by the class -1, with a representativeness of 22%. It means that most of the terms present in this lexicon are slightly positive or slightly negative.

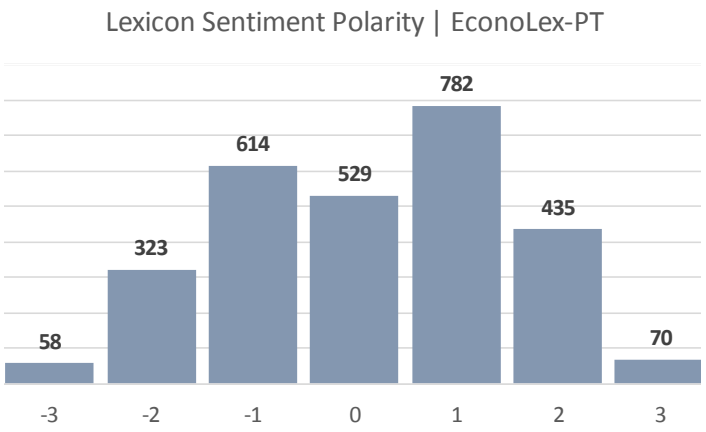

Figure 4 Sentiment polarity histogram for EconoLex-PT Lexicon

4.2 Sentiment Analysis according to the Human Classification

The domain experts from the area of linguistics have evaluated 45 articles of opinion from the economic domain and analyzed the contribution of each phrase from the point of view of text understanding and characterization of the overall sentiment expressed. They have decided to maintain 370 phrases and then manually assigned a sentiment score in a range from -3 to +3 (see **Appendix B: Expert analysis on Text #2**).

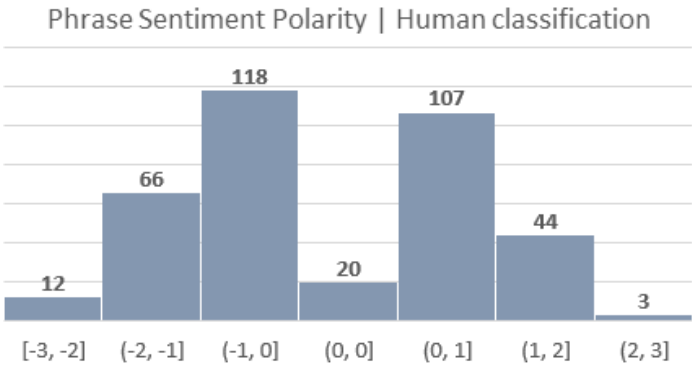

Figure 5 Sentiment polarity histogram according to the human classification (FLUP)

Figure 5 gives an overview of the scores assigned to each one of the 370 phrases evaluated by the researchers. It shows that most of the phrases are slightly negative or slightly positive, respectively with 118 and 107 cases, representing together a total of 61% of all phrases. Moreover, it can be noticed that overall there is more negative phrases than positive ones, 196 against 154 cases.

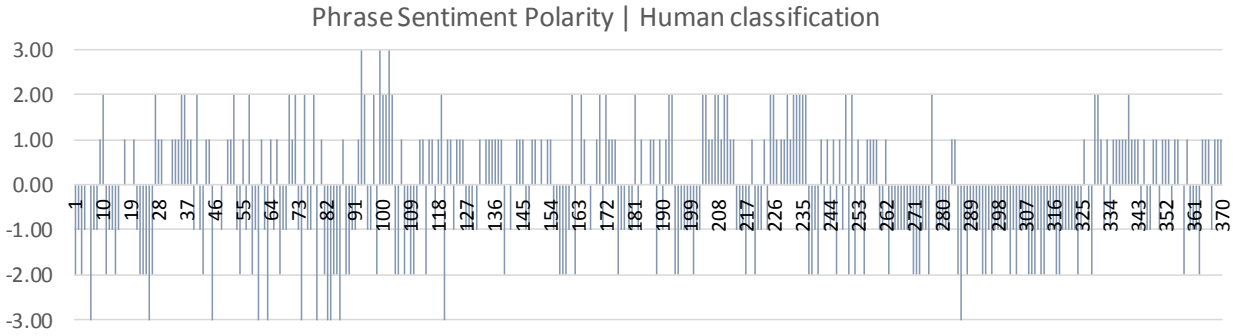

Figure 6 Sentiment polarity according to the human classification (FLUP)

Figure 5 gives an overview of the scores assigned to each one of the 370 phrases evaluated by the researchers.

4.3 Step 1: Initial SA

In the first step, the sentiment polarity of each phrase was calculated by summing up the sentiment values of all terms that occur simultaneously in the phrase and in the lexicon. Then, the sentiment value existing in the lexicon is assigned to the correspondent term. This procedure was performed for each lexicon independently and the distribution of sentiment polarity is presented in the Figure **7 Phrase Sentiment polarity comparison in the Step 1**.

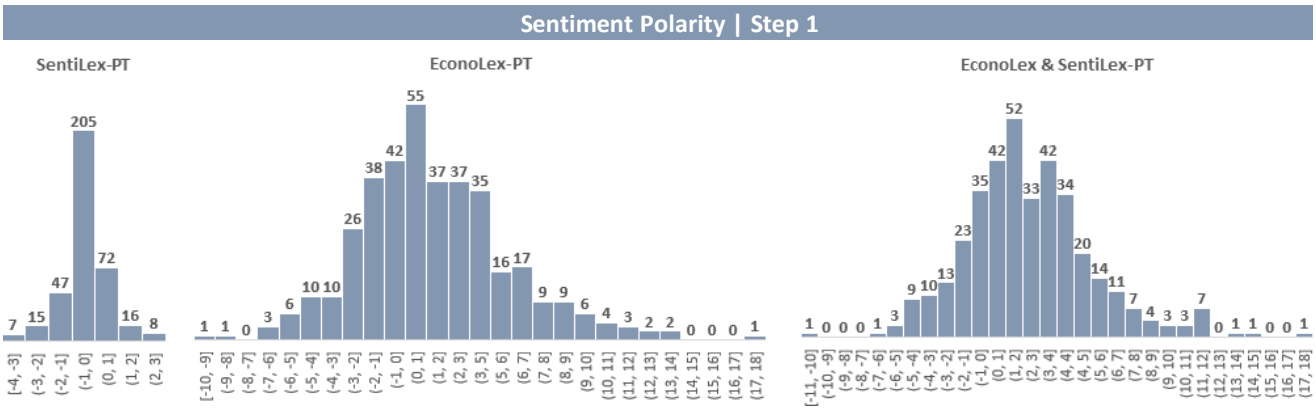

Figure 7 Phrase Sentiment polarity comparison in the Step 1

For illustration, let us analyze one example, namely the phrase:

"Por mais avanços que a tecnocracia europeia se mostre disposta a dar, na sequência da crise do euro, a predominância de um pensamento conservador nas esferas de poder torna previsível que fossem impostas de novo medidas violentas de contenção orçamental." (extracted from the text Até que uma crise nos separe, written by João Madeira and published by Jornal Económico in 30/11/2017).

The human subject assigned a value of -2 to this the phrase, indicating that this phrase presents a negative sentiment.

The SA system based on SentiLex, attributed the score of -1. The only term found in this lexicon was *"impostas"* [-1].

The SA system based on EconoLex attributed the score of -3. The terms found in this lexicon were *"contenção orçamental"* [-1], *"medidas violentas"* [-2], *"pensamento conservador"* [-1], *"crise"* [-2], *"europeia"* [0], *"mais"* [+2] and *"novo"* [+1].

Based on both EconoLex & SentiLex, the same phrase had a score of -4. The terms found in this lexicon were a union of the terms above. Should the same term appear in both lexica (which is not the case here), preference is given to the term appearing in EconoLex.

We note that the range of values obtained with the SA system based on EconoLex & SentiLex-PT differs a lot from the values provided by human subject. If we were to calculate the error measure MAE, it would be rather high. This observation led us to introduce scale adjustment, which is discussed in the next section.

4.4 Step 2: SA after applying the Scale Adjustment Factor

In the second step, a factor was calculated to adjust the difference of scales between the predicted sentiment value and the true sentiment value. The factor is calculated by dividing the standard deviation of the predicted values (σ_{pred}) by the standard deviation of the true value (σ_{true}), resulting in one factor for each lexicon.

In our case σ_{pred} was 1.02 for SentiLex, 3.65 for EconoLex and 3.85 for EconoLex & Sentilex; and σ_{true} was 1.45. Therefore, the adjustment factor was 0.70 for SentiLex, 2.51 for EconoLex and 2.65 for EconoLex & Sentilex.

The initial polarity sentiment value is then divided by this factor and results in a new and rescaled sentiment value.

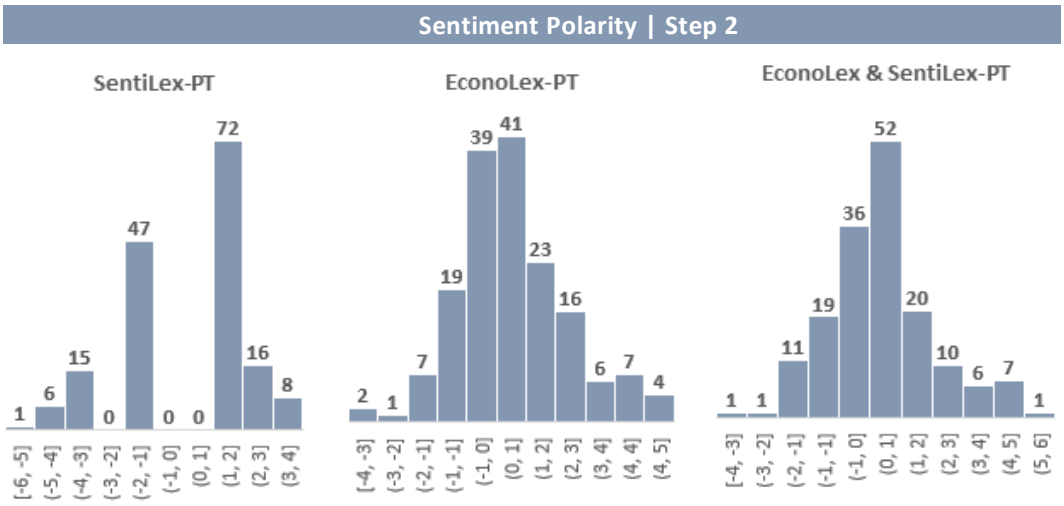

Figure 8 Phrase Sentiment Polarity Comparison in the Step 2

After applying the adjustment factor described above, the sentiment polarity was rescaled and the histogram width was significantly decreased (see Figure 8) when comparing to the previous data (see **Figure 7**).

4.5 Step 3: SA after performing Cross Validation

In the third step, Leave One Out Cross Validation (LOOCV) was used to evaluate the accuracy of the predictive model.

The procedure was performed on 44 of the 45 texts and then the performance of the algorithm is tested on the 45th. Then, the 45th text represents the text on which our SA system is tested, while the other 44 texts are used to extract the domain-specific terms. This procedure is repeated 45 times, each time leaving out a different text, which is used for testing.

It is important to stress in each cycle of LOOCV the lexicon was adjusted as described to ensure that we have unbiased evaluation.

The results show that both EconoLex-PT and EconoLex & SentiLex-PT achieve in this step their best performance, as can be seen in the **Figure 9**, in which the respective histograms of Phrase Sentiment Polarity present their narrowest version.

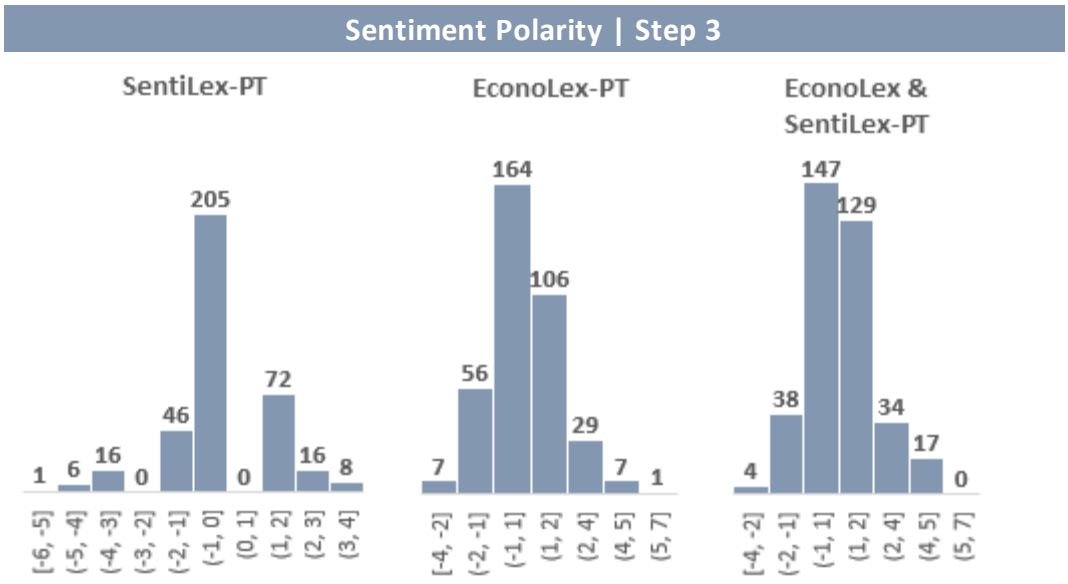

Figure 9 Phrase Sentiment Polarity Comparison in the Step 3

4.6 Comparing Results

In the first step, in which the sentiment polarity is calculated by the usual method, SentiLex presented the best performance between the three lexica. However, that result only happened due to a difference of scales between the predicted sentiment value and the sentiment value attributed by domain experts. This difference led to a significant negative impact in the specific-domain versions.

In the second step, in which the adjustment factor was applied to the data, the sentiment polarity was rescaled and those inconsistencies were eliminated. Hence, EconoLex as well as EconoLex & SentiLex presented a reduction in the error rate versus the real value and also an improvement in their performance.

In the third and last step, the accuracy of the predictive model was assessed and the lexica was adjusted to ensure that the evaluation was not biased.

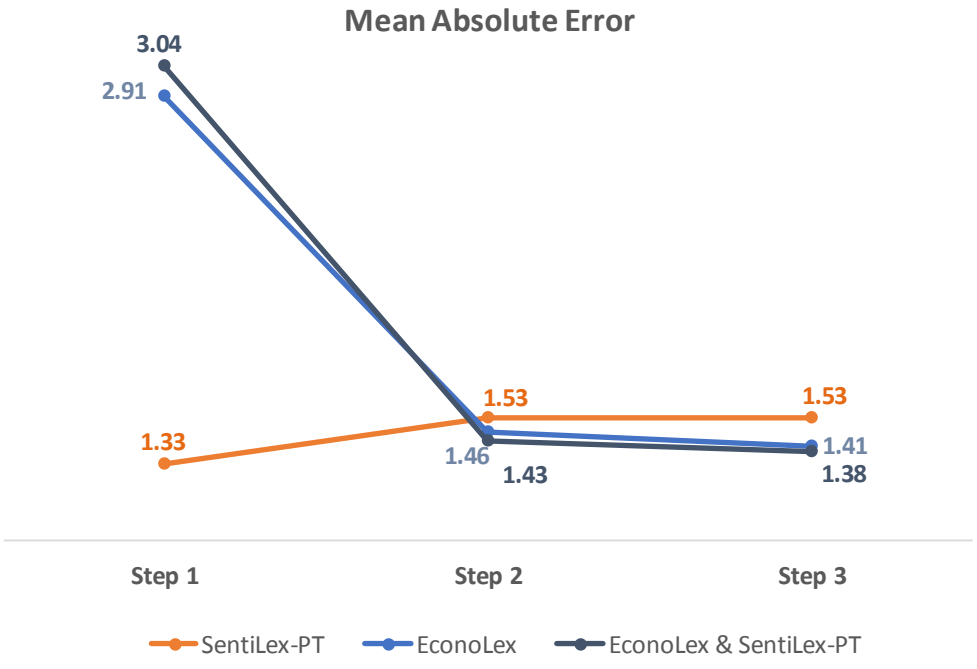

Figure 10 MAE evolution

Mean Absolute Error (MAE) was the evaluation measure chosen for this work due to this capability to capture the extension of the errors, while comparing the predicted value for each lexicon against the actual value.

Figure 10 illustrates MAE evolution in each step and it also reveals that the best results were achieved by the EconoLex & SentiLex version since this lexicon presented the lowest MAE at the end of the analysis.

- **Wilcoxon Test for statistical significance**

The Wilcoxon non-parametric test with a standard setting of 95% for the confidence interval was performed in R to assess the statistical significance of our results.

This non-parametric test ranked the absolute values of differences between the paired observations for each version of the SA system which used a different lexicon and calculated a statistic on the number of negative and positive differences.

The resulting p-value was 0.09874 for EconoLex & SentiLex versus SentiLex on it own. A similar test was performed also for EconoLex & SentiLex versus SentiLex. The p-value was higher in this case.

Since the p-value was higher than 0.05 in both cases, despite being relatively close in the first scenario, the conclusion is that the differences between the paired observations are not statistically significantly different.

5. Conclusions and Future work

The goal of this thesis was to verify whether the combination of EconoLex-PT and SentiLex-PT achieves better performance than just SentiLex-PT on its own, analysing the contribution that a domain-specific lexicon plays in a sentiment analysis system.

Initially, the punctuation of each phrase was defined by the sum of polarity of its terms based on the values in the lexicon. Then, a factor was applied to perform an adjustment needed due to a difference of scales between the predicted values and the actual values. Finally, the cross-validation technique was performed to evaluate the accuracy of the predictive model. The results of each step above are shown in the **Table 3.**

	Step 1	Step 2	Step 3
	Initial data	Data after applying the Scala adjustment factor	Data after performing Leave One Out Cross Validation
Lexicon	MAE	MAE	MAE
SentiLex-PT	1.33	1.53	1.53
EconoLex	2.91	1.46	1.41
EconoLex & SentiLex-PT	3.04	1.43	1.38

Table 3 MAE evolution

In fact, the Men Absolute Error calculations shown in the **Table 3** lead us to conclude that a domain-specific lexicon achieves better performance than a general-purpose lexicon. However, the Wilcoxon non-parametric result in a p-value of 0.09874 for SentiLex-PT on it owns and EconoLex & SentiLex-PT and a p-value of 0.1856 for SentiLex-PT and EconoLex-PT. Therefore, the non-parametric test indicates that the differences of MAE between the domain-specific lexicon and the general-purpose lexicon were not statistically significant.

The main conclusion is that, for observing a marked increase of performance on new texts, it is necessary to have a more robust corpus since the 45 texts analyzed were not enough for creating a robust lexicon and generating a sufficient number of repeated terms across texts.

For future works, the suggestions are the following:

- Improve the lexicon by including new entries;
- Improve the lexicon with synonyms;
- Improve linguistic analysis.

6. References

Almatarneh, S. and Gamallo, P. (2017). Automatic construction of domain-specific sentiment lexicons for polarity classification. In: De la Prieta, F. et al. (eds), Advances in Intelligent Systems and Computing. Porto, Portugal.

Balakrishnan, V. and Yemoh-Lloyd, E. (2014). Stemming and Lemmatization: A Comparison of Retrieval Performances. In: Lecture Notes on Software Engineering, Vol. 2, No. 3.

Batrinca, B. and Treleaven, P. C. (2014). Social media analytics: a survey of techniques, tools and platforms. AI & SOCIETY, pages 1–28.

Boiy, E and Moens, M. F. (2009). A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts. In: Proceeding ICAIL '07 Proceedings of the 11th International Conference on Artificial Intelligence and Law Pages 225-230.

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocation expressions in large textual databases. Proceedings of the RIAO, pp. 38-43.

Ding, Xiaowen; Liu, Bing and Yu, P. S. (2008). A Holistic Lexicon-Based Approach to Opinion Mining. In : WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining. Pages 231-240.

Edelman, B.; Ostrovsky, M. and Schwarz, M. (2007). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. The American Economic Review, 97(1): p. 242-259.

Esuli, A and Sebastiani, F. (2006). SentiWordNet: a publicly available lexical resource for opinion mining. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 5), ed. N Calzolari, K Choukri, A Gangemi, B Maegaard, J Mariani, et al., pp. 417–22. Genoa, Italy: Eur. Lang. Resour. Assoc.

Farzindar, A. and Inkpen, D. (2018). Natural Language Processing for Social Media. In: Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers Graeme Hirst, Series Editor.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Feldman, R. and J. Sanger (2007). Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.

Forte, A. C. and Brazdil, P.B. (2016). Determining the Level of Clients' Dissatisfaction from Their Commentaries. In: Silva J., Ribeiro R., Quaresma P., Adami A., Branco A. (eds) Computational Processing of the Portuguese Language. PROPOR 2016. Lecture Notes in Computer Science, vol. 9727.

Gaikward, S. V.; Chaugule, Archama and Patil, Pramod (2014). Text Mining Methods and Techniques. In: International Journal of Computer Applications (0975 – 8887).

Goldberg, Y. (2017). Neural Network Methods in Natural Language Processing in Synthesis Lectures on Human Language Technologies. Graeme Hirst, University of Toronto.

Gouws, R. H.; Heid, U.; Schweickard, W. and Wiegand, H. E. (2013). Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography. De Gruyter Mouton.

Hu, M. and Liu, B. (2004). Mining and Summarizing Customer Reviews. In: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 168–177. ACM.

Huang, Z., Zeng, D.D. and Chen, H. (2007). Analyzing Consumer-Product Graphs: Empirical Findings and Applications in Recommender Systems.Management Science, 53(7): p. 1146-1164.

Jivani, A. G. (2011). A Comparative Study of Stemming Algorithms. In: Int. J. Comp. Tech. Appl., Vol 2 (6), 1930-1938.

Liu, B. (2011). Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. 2nd ed. Chicago: Springer.

Liu, B. (2010). Sentiment Analysis and Subjectivity. Handbook of natural language processing, 2:627–666. 2nd Edition, CRC Press, New York.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. G. Hirst, ed., Morgan & Clayopool Publishers.

Liu, B. and Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis in Mining Text Data, pages 415–463. Springer.

Makrehchi, M. and Kamel, M.S. (2008). Automatic Extraction of Domain-Specific Stopwords from Labeled Documents. In: Macdonald C., Ounis I., Plachouras V., Ruthven I., White R.W. (eds) Advances in Information Retrieval. ECIR 2008. Lecture Notes in Computer Science, vol 4956. Springer.

Manning, C. D.; Raghavan, P. and Schutze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

Medhat, W.; Hassan, A. and Korashy, H. (2014). Electrical Engineering: Sentiment Analysis Algorithms and Applications: A survey. Ain Shams Engineering Journal. 5, 1093-1113. ISSN: 2090-4479.

Pang, B.; Lee, L. and Vaithyanathan, S. (2002). Thumbs up?: Sentiment Classification Using Machine Learning Techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pages 79–86. Association for Computational Linguistics.

Schone, P. and Jurafsky, D. (2001). Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem?

Silva, Mário J.; Carvalho, Paula and Sarmento, Luís (2012). Building a Sentiment Lexicon for Social Judgement Mining". Lecture Notes in Computer Science (LNCS), International Conference on Computational Processing of the Portuguese Language (PROPOR), Springer, pp. 218-228.

Taboada, M.; Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011). Lexicon based methods for sentiment analysis. Computational linguistics, 37(2):267–307.

Turney, P. D. (2002). Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 417–424. Association for Computational Linguistics

Wiebe, J. M.; Bruce, R. F. and O'Hara, T. P. (1999). Development and Use of a Gold-Standard Data Set for Subjectivity Classifications. In: Proceedings of the Association for Computational Linguistics (AC L-1999).

Wilson, T.; Wiebe, J. and Hwa, Rebecca (2004). Just How Mad are You? Finding Strong and Weak Opinion Clauses. In Proceedings of National Conference on Artificial Intelligence (AAAI-2004).

Appendix A: Corpus summary

Corpus (1/2)

Text	Article	Date	Source	Author
1	Crescimento da Amazon é um "presente envenenado" para o mercado de trabalho	05/12/2017	Jornal Expresso	Jornal Expresso
2	Até que uma crise nos separe	30/11/2017	Jornal Económico	João Madeira
3	Portugal foi o segundo país mais beneficiado com as compras de dívida do BCE	07/12/2017	Jornal Económico	João Madeira
4	Europa, a hipócrita	19/07/2017	Jornal Expresso	Marco Capitão Ferreira
5	Eurostat: Economias da zona euro e UE crescem 2,6% no 3.º trimestre	07/12/2017	Jornal Económico	Jornal Económico com Lusa
6	Consórcio da Galp assegura 3,9 mil milhões de euros para Moçambique	06/12/2017	Jornal Expresso	Miguel Prado
7	Geringonça a gripar	30/11/2017	Jornal Económico	Francisco Proença de Carvalho
8	IRS: manusear com cautela	13/09/2017	Jornal Expresso	Marco Capitão Ferreira
9	Declínio	29/11/2014	Jornal Expresso	Luís Todo Bom
10	Vida para além do défice	06/12/2017	Jornal Expresso	Marco Capitão Ferreira
11	O outro lado da banca	07/11/2017	Observador	José Paulo Miller
12	Temos-Orçamento	29/11/2017	Expresso	Marco Capitão Ferreira
13	Orçamento do estado – Um olhar sobre o IRS!	25/10/2017	Jornal Público	Ana Duarte
14	Uma nova visão para o mundo	03/11/2017	Visão	Pedro Camacho
15	Curva (demasiado) apertada	10/11/2017	Visão	Paulo M. Santos
16	Depois do brexit, a economia das "rendas" do Reino Unido deixará de ser viável	20/11/2017	Diário de Notícias	Wolfgang Münchau
17	Exame ao governo	28/09/2014	Visão	João Paulo Vieira, Carla Alves Ribeiro
18	Todos os caminhos vão dar ao défice	19/09/2013	Visão	Clara Teixeira
19	O amigo oculto	28/06/2007	Visão	Paulo M. Santos
20	Um Orçamento que consolida a alternativa	27/11/2017	Expresso	João Galamba
21	Um Orçamento de costas para as empresas	27/11/2017	Expresso	Diogo Agostinho
22	Economia Global em Divergência	06/11/2017	Público	Pedro Jordão
23	Economia portuguesa chega pouco produtiva ao pós-crise	10/12/2017	Jornal de Negócios	Nuno Aguiar
24	Dignificar o trabalho para uma economia com futuro	01/05/2017	Diário de Notícias	António Costa
25	O que diz, na realidade, a teoria económica sobre a austeridade	14/05/2012	Jornal Público	Armando Pires

Corpus (2/2)

Text	Article	Date	Source	Author
26	Infarmed entre o norte e o desnorte	25/11/2017	Jornal Económico	António Freitas de Sousa
27	Ciclos de consequências	24/11/2017	Jornal Económico	Safaa Dib
28	Dignificar o trabalho para uma economia com futuro	01/05/2017	Jornal Económico	António Costa
29	Paralisia extrema	23/11/2017	Jornal Económico	Ricardo Leite Pinto
30	Agravar a carga tributária afasta o investimento	17/11/2017	Jornal Económico	Jorge Jordão
31	Poderá o crédito habitação ser um bom instrumento de poupança?	10/11/2017	Jornal Económico	Gustavo Soares
32	A chama invisível	10/11/2017	Jornal Económico	Luís Tavares Bravo
33	Dois anos de bons resultados	25/11/2017	Jornal de Notícias	Pedro Silva Pereira
34	Desemprego abaixo dos 9%	12/08/2017	Jornal Económico	Pedro Silva Pereira
35	Mercosul: uma oportunidade	24/11/2017	Jornal Económico	João Gonçalves Pereira
36	A actual proposta de Orçamento do Estado não promove uma economia de sucesso para Portugal	09/11/2017	Observador	Alexandre Patrício Gouveia
37	A falha do Orçamento do Estado para 2018	02/11/2017	Diário de Notícias	Pedro Filipe Soares
38	Tácticas de austeridade escondida	26/10/2017	Observador	Helena Garrido
39	Um Orçamento à esquerda	17/10/2017	Observador	Joana Vicente
40	Pequenas notas sem importância	01/11/2017	Observador	Maria João Avillez
41	Da vergonha perdida e por perder	12/11/2017	Observador	Helena Matos
42	O IRS no Orçamento do Estado – Boas notícias, mas não para todos!	17/10/2017	Negócios	Ana Duarte
43	Opinião: O agridoce do Orçamento do Estado para 2018	15/10/2017	Dinheiro Vivo	Carlos Lobo
44	Primeiras Impressões sobre o Orçamento do Estado para 2018	14/10/2017	TSF	Carlos Baptista Lobo
45	Opinião. Falemos de Orçamento de Estado	15/07/2017	Dinheiro Vivo	António Saraiva

Appendix B: Expert analysis on Text #2

Relevant sentence	Sentence Polarity	Noun Polarity	Adjective Polarity	Relevant expression Polarity
Os **sucessivos casos** que envolvem o Governo de António Costa – da gestão dos incêndios à reviravolta na votação da taxa sobre as rendas energéticas – banalizaram mais uma vez os vaticínios de que a morte da 'geringonça' estará a caminho.	-1	Casos: 0	Sucessivos: 0	Sucessivos casos: 0
Naqueles idos do fim de socratismo, com Passos na oposição, o Bloco teve a "bravata" de ser o primeiro partido a apresentar uma moção de censura contra o **Governo minoritário** de Sócrates.	0	Governo: 0	Minoritário: -1	Governo minoritário: -1
Face às eleições anteriores, o Bloco perdeu metade dos votos, metade dos deputados e entrou numa **crise profunda**.	-3	Crise: -2	Profunda: -2	Crise profunda: -3
(...) só um brilharete de Mariana Mortágua na comissão do BES resgatou o partido de um dos períodos de maior declínio da sua **breve história**.	-1	História: 0	Breve: 0	Breve história: 0
O eleitorado de esquerda podia não gostar de Sócrates, mas não gostava ainda mais de um PSD com um **projecto** político e económico **ultra-liberal**, como veio mais tarde a comprovar-se.	-1	Projecto: 0	Ultra-liberal: 0	Projeto ultra-liberal: 0
Qualquer partido funciona com **incentivos eleitorais** e o Bloco não é exceção (...)	1	Incentivos: 1	Eleitorais: 0	Incentivos eleitorais: 1
(...) parece ter aprendido uma **valiosa lição** com a condução dos trabalhos na queda de Sócrates.	2	Lição: 1	Valiosa: 2	Valiosa lição: 2
Não havendo uma **falha grave** que ponha em causa a credibilidade do partido, ou uma queda nas sondagens, o partido deverá continuar a apoiar o Governo.	-2	Falha: -1	Grave: -2	Falha grave: -3
O real entrave a uma continuidade do acordo seria uma crise que empurrasse de novo o país para um período recessivo (...)	-1	Entrave: -1 Período: 0	Real: 1 Recessivo: -1	Real entrave: -1 Período recessivo: -1
(...) com a inerente **consequência negativa** nas contas públicas – à boleia da queda das receitas fiscais e do aumento dos custos com o desemprego.	-1	Consequência: 0	Negativa: -1	Consequência negativa: -1
Por mais avanços que a tecnocracia europeia se mostre disposta a dar, na sequência da crise do euro, a predominância de um **pensamento conservador** nas esferas de poder torna previsível que fossem impostas de novo **medidas violentas** de **contenção orçamental**.	-2	Pensamento: 0 Medidas: 0 Contenção: -1	Conservador: 0 Violentas: -1 Orçamental: 0	Pensamento conservador: -1 Medidas violentas: -2 Contenção orçamental: -1
O **delicado compromisso** que o PS tem conseguido manter, entre o cumprimento das metas orçamentais e o cepticismo do Bloco e do PCP face às normas comunitárias no que diz respeito às políticas económicas e orçamentais, entraria em colapso.	-1	Compromisso: 0	Delicado: 1	Delicado compromisso: -1

Appendix C: Algorithm

```
library(tidyverse)
library(tokenizers)
library(tm)
library(dplyr)
library(xlsx)
library(plyr)

text <- readLines("Texts.txt")
text

lexicon_base <- read.csv("cv_SentiLex-flex-PT02_all.csv", sep=";", header=FALSE)
names(lexicon_base) <- c("Term", "Value", "Source")
lexicon_base

Econolex_base <- read.csv("cv_EconoLex-flex_all.csv", sep=";", header=FALSE)
names(Econolex_base) <- c("Term", "Value", "Source")
Econolex_base

text_experts <- read.csv("Texts_FLUP.csv", sep=";", header=FALSE)
text_experts[,1] <- tolower(text_experts[,1])
names(text_experts) <- c("FLUP_Phrase", "FLUP_Value")

# General function
analysis <- function(texts, lexicons_base, Econolexs_base, texts_experts, NGRAMS, KWORDS){
a <- 0
phrase.value <- data.frame()
n.eco <- data.frame()
nEconolex.uni <- filter(Econolexs_base, !grepl( " ", Econolexs_base$'Term'))
nEconolex.multi <- filter(Econolexs_base, grepl( " ", Econolexs_base$'Term'))

for (i in 1:length(texts)) {
    treino <- texts[-i]

        for (q in 1: length(treino)) {
            pre.processed.phrase1 <- list()
            b <- tokenize_sentences(treino[[q]])
            all.phrases <- data.frame(b)
            n.eco.phrases <- data.frame()

            for (j in 1:sum(table(all.phrases))) {
                qq <- data.frame()
                qqdf1 <- data.frame()
                qqdf11 <- data.frame()

                #Econolex multipalavra
                pre.processed.phrase <- tolower(all.phrases[j,])
                multiwords <- tokenize_skip_ngrams(pre.processed.phrase, n = NGRAMS, k = KWORDS)
                tab1 <- table(multiwords[[1]])
                multiwords.df <- data_frame(word = names(tab1), count = as.numeric(tab1))
                multiwords.values.df1 <- left_join(multiwords.df, nEconolex.multi %>% select(Term, Value), by = c("word" = "Term"))
                x <- filter(multiwords.values.df1, multiwords.values.df1[,3] != "NA")
                multiwords.values.df1 <- unique(x)
                sentence.value11 <- cbind(multiwords.values.df1, multiwords.values.df1[2]*multiwords.values.df1[3])
                names(sentence.value11) <- c('Word', 'Freq', 'Value', 'Total.Value')

                if (nrow(multiwords.values.df1)!=0) {
                for (q in 1: nrow(multiwords.values.df1)){
                qq <- table(tokenize_words(multiwords.values.df1[q,1][[1]]))
                qqdf1 <- data_frame(word = names(qq), count = as.numeric(qq))
                qqdf11 <- aggregate(. ~ word, rbind(qqdf1, qqdf11), FUN=sum)
                }} else { qqdf11 <- c()}

                pre.processed.phrase1[[j]] <- removeWords(pre.processed.phrase, c(qqdf11[,1]))

                #Econolex unipalavra
                words1 <- tokenize_words(pre.processed.phrase1[[j]])
                tab21 <- table(words1[[1]])
                words.df1 <- data_frame(word = names(tab21), count = as.numeric(tab21))

                words.values.df1 <- left_join(words.df1, nEconolex.uni %>% select(Term, Value), by = c("word" = "Term"))
```

```
            m <- filter(words.values.df1, words.values.df1[,3] != "NA")
            words.values.df1 <- unique(m)
            sentence.value12 <- cbind(words.values.df1, words.values.df1[2]*words.values.df1[3])
            names(sentence.value12) <- c('Word', 'Freq', 'Value', 'Total.Value')

            sentence.value1 <- bind_rows(sentence.value11, sentence.value12)
            names(sentence.value1) <- c('Word', 'Freq', 'Value', 'Total.Value')
            n.lx <- sentence.value1[,c(1,3)]

            n.eco.phrases <- bind_rows(n.eco.phrases, n.lx)

        }
        names (n.eco.phrases) <- c("Term", "Value")
        n.eco <- bind_rows(n.eco, n.eco.phrases)
    }

n.eco[,3] <- c("Econo.Selected.Texts")
names(n.eco) <- c("Term", "Value", "Source")
new.econolex <- unique(n.eco)
new.lexicon <- bind_rows(n.eco, lexicons_base)
names(new.lexicon) <- c("Term", "Value", "Source")

Sentilex.uni <- filter(lexicons_base, !grepl( " ", lexicons_base$'Term'))
Sentilex.multi <- filter(lexicons_base, grepl( " ", lexicons_base$'Term'))
Econolex.uni <- filter(new.econolex, !grepl( " ", new.econolex$'Term'))
Econolex.multi <- filter(new.econolex, grepl( " ", new.econolex$'Term'))
Senti.Econo.uni <- filter(new.lexicon, !grepl( " ", new.lexicon$'Term'))
Senti.Econo.multi <- filter(new.lexicon, grepl( " ", new.lexicon$'Term'))



# Evaluation
pre.processed.phrase1 <- list()
pre.processed.phrase2 <- list()
pre.processed.phrase3 <- list()

b <- tokenize_sentences(texts[[i]])
all.phrases <- data.frame(b)

for (j in 1:sum(table(all.phrases))) {
    qq <- data.frame()
    qqdf1 <- data.frame()
    qqdf11 <- data.frame()
    qqdf2 <- data.frame()
    qqdf21 <- data.frame()
    qqdf3 <- data.frame()
    qqdf31 <- data.frame()

    #avaliação pelos lexicon multipalavra
    pre.processed.phrase <- tolower(all.phrases[j,])
    multiwords <- tokenize_skip_ngrams(pre.processed.phrase, n = NGRAMS, k = KWORDS)
    tab1 <- table(multiwords[[1]])
    multiwords.df <- data_frame(word = names(tab1), count = as.numeric(tab1))
    multiwords.values.df1 <- left_join(multiwords.df, Sentilex.multi %>% select(Term, Value), by = c("word" = "Term"))
    x <- filter(multiwords.values.df1, multiwords.values.df1[,3] != "NA")
    multiwords.values.df1 <- unique(x)
    multiwords.values.df2 <- left_join(multiwords.df, Econolex.multi %>% select(Term, Value), by = c("word" = "Term"))
    w <- filter(multiwords.values.df2, multiwords.values.df2[,3] != "NA")
    multiwords.values.df2 <- unique(w)
    multiwords.values.df3 <- left_join(multiwords.df, Senti.Econo.multi %>% select(Term, Value), by = c("word" = "Term"))
    z <- filter(multiwords.values.df3, multiwords.values.df3[,3] != "NA")
    multiwords.values.df3 <- unique(z)

    sentence.value11 <- cbind(multiwords.values.df1, multiwords.values.df1[2]*multiwords.values.df1[3])
    sentence.value21 <- cbind(multiwords.values.df2, multiwords.values.df2[2]*multiwords.values.df2[3])
    sentence.value31 <- cbind(multiwords.values.df3, multiwords.values.df3[2]*multiwords.values.df3[3])
    names(sentence.value11) <- c('Word', 'Freq', 'Value', 'Total.Value')
    names(sentence.value21) <- c('Word', 'Freq', 'Value', 'Total.Value')
    names(sentence.value31) <- c('Word', 'Freq', 'Value', 'Total.Value')

```

```
            if (nrow(multiwords.values.df1)!=0) {
            for (q in 1: nrow(multiwords.values.df1)){
            qq <- table(tokenize_words(multiwords.values.df1[q,1][[1]]))
            qqdf1 <- data_frame(word = names(qq), count = as.numeric(qq))
            qqdf11 <- aggregate(. ~ word, rbind(qqdf1, qqdf11), FUN=sum)
            }} else { qqdf11 <- c()}


            if (nrow(multiwords.values.df2)!=0) {
            for (q in 1: nrow(multiwords.values.df2)){
            qq <- table(tokenize_words(multiwords.values.df2[q,1][[1]]))
            qqdf2 <- data_frame(word = names(qq), count = as.numeric(qq))
            qqdf21 <- aggregate(. ~ word, rbind(qqdf2, qqdf21), FUN=sum)
            }} else { qqdf21 <- c()}

            if (nrow(multiwords.values.df3)!=0) {
            for (q in 1: nrow(multiwords.values.df3)){
            qq <- table(tokenize_words(multiwords.values.df3[q,1][[1]]))
            qqdf3 <- data_frame(word = names(qq), count = as.numeric(qq))
            qqdf31 <- aggregate(. ~ word, rbind(qqdf3, qqdf31), FUN=sum)
            }} else { qqdf31 <- c()}

            pre.processed.phrase1[[j]] <- removeWords(pre.processed.phrase, c(qqdf11[,1]))
            pre.processed.phrase2[[j]] <- removeWords(pre.processed.phrase, c(qqdf21[,1]))
            pre.processed.phrase3[[j]] <- removeWords(pre.processed.phrase, c(qqdf31[,1]))

        words1 <- tokenize_words(pre.processed.phrase1[[j]])
        words2 <- tokenize_words(pre.processed.phrase2[[j]])
        words3 <- tokenize_words(pre.processed.phrase3[[j]])
        tab21 <- table(words1[[1]])
        tab22 <- table(words2[[1]])
        tab23 <- table(words3[[1]])
        words.df1 <- data_frame(word = names(tab21), count = as.numeric(tab21))
        words.df2 <- data_frame(word = names(tab22), count = as.numeric(tab22))
        words.df3 <- data_frame(word = names(tab23), count = as.numeric(tab23))
        words.values.df1 <- left_join(words.df1, Sentilex.uni %>% select(Term, Value), by = c("word" = "Term"))
        m <- filter(words.values.df1, words.values.df1[,3] != "NA")
        words.values.df1 <- unique(m)
        words.values.df2 <- left_join(words.df2, Econolex.uni %>% select(Term, Value), by = c("word" = "Term"))
        n <- filter(words.values.df2, words.values.df2[,3] != "NA")
        words.values.df2 <- unique(n)
        words.values.df3 <- left_join(words.df3, Senti.Econo.uni %>% select(Term, Value), by = c("word" = "Term"))
        o <- filter(words.values.df3, words.values.df3[,3] != "NA")
        words.values.df3 <- unique(o)
        sentence.value12 <- cbind(words.values.df1, words.values.df1[2]*words.values.df1[3])
        sentence.value22 <- cbind(words.values.df2, words.values.df2[2]*words.values.df2[3])
        sentence.value32 <- cbind(words.values.df3, words.values.df3[2]*words.values.df3[3])
        names(sentence.value12) <- c('Word', 'Freq', 'Value', 'Total.Value')
        names(sentence.value22) <- c('Word', 'Freq', 'Value', 'Total.Value')
        names(sentence.value32) <- c('Word', 'Freq', 'Value', 'Total.Value')
        sentence.value1 <- bind_rows(sentence.value11, sentence.value12)
        sentence.value2 <- bind_rows(sentence.value21, sentence.value22)
        sentence.value3 <- bind_rows(sentence.value31, sentence.value32)
        names(sentence.value1) <- c('Word', 'Freq', 'Value', 'Total.Value')
        names(sentence.value2) <- c('Word', 'Freq', 'Value', 'Total.Value')
        names(sentence.value3) <- c('Word', 'Freq', 'Value', 'Total.Value')
        a<-nrow(phrase.value)
        phrase.value[a+1,1]<- i
        phrase.value[a+1,2]<- j
        phrase.value[a+1,3]<- b[[1]][[j]]
        phrase.value[a+1,4]<- sum(filter(sentence.value1, sentence.value1$'Total.Value' != 'NA')$ 'Total.Value')
        phrase.value[a+1,5]<- sum(filter(sentence.value2, sentence.value2$'Total.Value' != 'NA')$ 'Total.Value')
        phrase.value[a+1,6]<- sum(filter(sentence.value3, sentence.value3$'Total.Value' != 'NA')$ 'Total.Value')
        phrase.value[a+1,7]<- paste(cbind(multiwords.values.df1[,1]))
        phrase.value[a+1,8]<- paste(cbind(words.values.df1[,1]))
        phrase.value[a+1,9]<- paste(cbind(multiwords.values.df2[,1]))
        phrase.value[a+1,10]<- paste(cbind(words.values.df2[,1]))
        phrase.value[a+1,11]<- paste(cbind(multiwords.values.df3[,1]))
        phrase.value[a+1,12]<- paste(cbind(words.values.df3[,1]))
    }
```

```
211
212     for (k in 1:nrow(phrase.value)){
213         if (tolower(phrase.value$V3[k]) %in% texts_experts[,1] == TRUE) {
214             phrase.value$V13[k] <- match(tolower(phrase.value$V3[k]),texts_experts[,1])
215             phrase.value$V14[k] <- texts_experts[phrase.value$V13[k], 2]
216             phrase.value$V15[k] <- abs(as.numeric(phrase.value$V14[k]) - as.numeric(phrase.value[k,4]))
217             phrase.value$V16[k] <- abs(as.numeric(phrase.value$V14[k]) - as.numeric(phrase.value[k,5]))
218             phrase.value$V17[k] <- abs(as.numeric(phrase.value$V14[k]) - as.numeric(phrase.value[k,6]))
219
220         } else{
221             phrase.value$V13[k] <- "Not Found"
222             phrase.value$V14[k] <- "Not Found"
223             phrase.value$V15[k] <- c("-")
224             phrase.value$V16[k] <- c("-")
225             phrase.value$V17[k] <- c("-")
226         }
227     }
228 }
229
230 names(phrase.value) <- c("Text N.", "Phrase Number", "Phrase", "Sentilex_Value", "Econolex_Value2", "Senti_Econo_Value",
231 "Senti_Multiwords", "Senti_Uniwords", "Econo_Multiwords", "Econo_Uniwords", "Senti_Econo_Multiwords","Senti_Econo_Uniwords",
232 "FLUP_Phrase", "FLUP_Value", "MAE_Sentilex", "MAE_Econolex", "MAE_Senti_Econo")
233
234 write.xlsx(phrase.value, "cv_output.xlsx", sheetName = "Results")
235 return(phrase.value)
236 }
237
238 analysis(text, lexicon_base, Econolex_base, text_experts, 3, 2)
239
240
241 # Wilcoxon
242
243 w_SentiLex <- as.numeric(readLines("MAE_Sentilex.txt"))
244 w_SentiLex
245
246 w_EconoLex <- as.numeric(readLines("MAE_EconoLex.txt"))
247 w_EconoLex
248
249 w_EconoLex_SentiLex <- as.numeric(readLines("MAE_EconoLex_SentiLex.txt"))
250 w_EconoLex_SentiLex
251
252 wilcox.test_sc1(w_EconoLex, w_SentiLex, mu=0, paired=TRUE, alternative="less")
253 wilcox.test_sc2(w_EconoLex_SentiLex, w_SentiLex, mu=0, paired=TRUE, alternative="less")
```